Journal Title Here, 2022, 1–12
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

hicream: a flexible framework to identify significantly different regions in Hi-C data

Elise Jorge,[1,*] Toby Dylan Hocking,[2] Pierre Neuvial,[3] Nathalie Vialaneix[4] and Sylvain Foissac[1,*]

[1]GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet-Tolosan, France, [2]LASSO lab, Département d'informatique, Université de Sherbrooke, Sherbrooke, QC, J1K-2R1, Canada, [3]Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS UPS, 31062, Toulouse, France and [4]Université Fédérale de Toulouse, INRAE, MIAT, 31326 Castanet-Tolosan, France

*Corresponding author. elise.jorge@inrae.fr, sylvain.foissac@inrae.fr

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Motivation: The three-dimensional (3D) organization of the genome plays a central role in many biological processes, and its disruption can critically impair cellular function. Detecting such changes is therefore crucial and can be achieved through differential analysis of Hi-C data. By comparing Hi-C data across biological conditions, differential Hi-C analysis aims to identify statistically significant and biologically relevant changes in chromatin organization. However, most existing methods either target a single predefined type of structure (e.g., TADs, loops, or compartments), thereby limiting their ability to uncover novel or complex structural variations, or produce highly local and disconnected results that are difficult to interpret in terms of higher-order genome organization.

Results: We introduce **hicream**, a novel framework for differential Hi-C analysis that identifies regions of arbitrary shape, allowing to uncover new differential structures. By combining pixel-level differential analysis and data-driven clustering to define candidate regions of the Hi-C interaction matrix, our approach provides an interpretable measure of the differential signal for each region through a post hoc inference strategy. The resulting differential regions can be explored using an interactive visualization interface.

Availability and Implementation: **hicream** is available at `https://cran.r-project.org/package=hicream`

Key words: Hi-C, differential analysis, *post hoc* inference, statistical testing

1. Introduction

In eukaryotic cells, chromosomes are compacted within the nucleus and organized into three-dimensional (3D) structures at different levels, including chromosomal territories, chromatin compartments, topologically associating domains (TADs), and chromatin loops. This three-dimensional (3D) organization regulates gene expression and plays a crucial role in various biological processes such as cell division and differentiation [Oudelaar et al., 2020]. It has been

© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026

shown to be involved in cancers and congenital anomalies [Lupiáñez et al., 2015, Spielmann et al., 2018].

The 3D genome architecture can be profiled using Hi-C [Lieberman-Aiden et al., 2009], which quantifies interaction frequencies between genomic intervals ("bins") along the chromatin. Hi-C data are commonly represented as sparse, symmetric interaction matrices, where each entry records the number of detected interactions between bins i and j, where (i, j) is called a "pixel". By comparing such matrices across biological conditions, differential Hi-C analysis aims at identifying statistically significant and biologically relevant changes in chromatin organization [Gjoni et al., 2025].

A first class of approaches to differential analysis of Hi-C data focuses on structures and aims to identify differential regions of the Hi-C matrix corresponding to an *a priori* defined type of structure, such as chromatin loops [Lareau and Aryee, 2017], TADs [Hua et al., 2024, Mourad, 2022] or A/B compartments [Chakraborty et al., 2022, Maigné et al., 2025]. This strategy constrains the analysis to one type of structure at a time, requiring to use multiple tools to capture variations across different levels of chromatin organization. Moreover, it either relies on rigid, structure-specific shape assumptions for differential regions, or is limited to detecting differential boundary locations. This leaves aside the diversity of shapes that structural variations can produce: For instance, Cresswell and Dozmorov [2020] identify several types of TAD changes with distinct differential region shapes, such as a triangle for TAD strengthening and a stripe for a boundary shift. Directly identifying differential regions in a shape-agnostic manner is therefore more flexible and comprehensive. Last, analyzing each structural type independently does not naturally account for complex chromatin variations arising from the interplay of multiple structural elements across different scales of organization.

A second class of approaches, referred to as "pixel-based" differential analysis, focuses on identifying differential pixels individually [Jorge et al., 2025b]. For each pixel (i, j), a test is performed and a p-value is computed. A global multiple testing correction using the Benjamini-Hochberg (BH) procedure [Benjamini and Hochberg, 1995] is then applied across all pixels to control the False Discovery Rate (FDR), defined as the expected proportion of false positives among the rejected hypotheses. The main limitation of this approach lies in the disconnected nature of the results and its dependence on matrix resolution. Restricting the analysis to independent pixels is not well suited to detecting structural variations that extend beyond regions of an arbitrary pixel size. To address this limitation, the **diffHic** package [Lun and Smyth, 2015] introduced the **diffHic clustering** method, which aims to form clusters of significant pixels based on adjusted p-values. However, as noted in the package documentation, this procedure does not provide formal statistical guarantees, as it relies on a potentially biased estimate of the cluster-level FDR.

To overcome the limitations of these approaches, we introduce **hicream**, a data-driven framework for detecting genomic structural variations by identifying differential regions of arbitrary size and shape between groups of Hi-C matrices. **hicream** builds upon a pixel-based differential analysis followed by a clustering step. It then combines the resulting clusters and associated p-values with a *post hoc* statistical inference framework [Goeman and Solari, 2011] to estimate the proportion of differential pixels within each cluster and to select clusters with the highest estimated proportions. We evaluated **hicream**, compared it against the **diffHic clustering** approach, and applied it to two independent real datasets, identifying differential regions. Results show consistency with external data, including active CTCF binding sites from ChIP-seq experiments and enriched transcription factor binding sites (TFBS).

2. Materials and Methods

2.1. Differential analysis of Hi-C data

In this article, we address the comparison between two groups of $r = r_1 + r_2$ Hi-C matrices, which are $(p \times p)$-symmetric matrices, with nonnegative integer entries for p bins. These matrices are denoted $M^{g,l}$, where $g \in \{1, 2\}$ stands for the group to which the matrix belongs (also called the "condition") and $l \in \{1, \cdots, r_g\}$ corresponds to the l-th replicate of condition g. Note that these replicates are not assumed to be paired between conditions. The matrix entries are denoted by $m_{ij}^{g,l}$. They represent the interaction frequency (in terms of number of reads) between bins i and j for the l-th matrix of condition g. Note that we only consider intra-chromosomal interactions, *i.e*, matrices corresponding to a single chromosome.

In this setting, the objective of Hi-C data differential analysis is to identify changes within the

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026

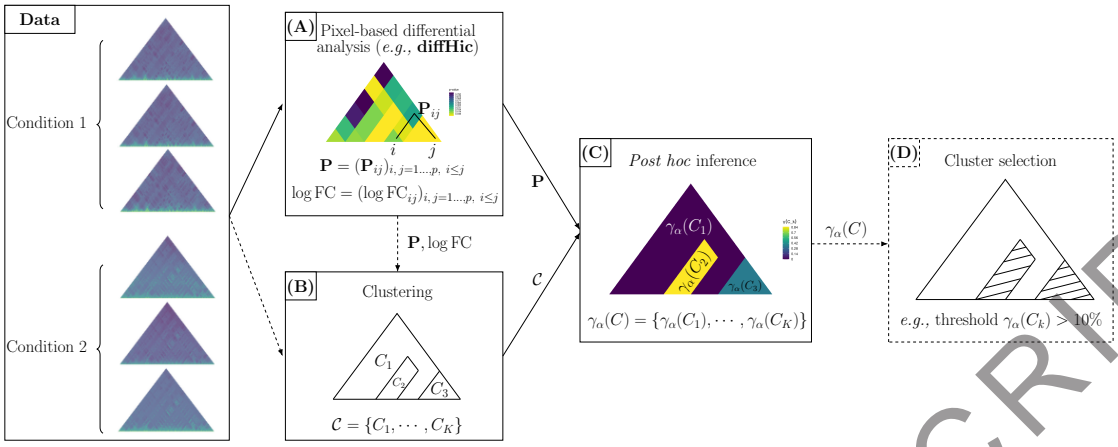


Fig. 1: **The hicream workflow.** Input data correspond to Hi-C matrices from two different biological conditions, with several replicates in each. (A) A pixel-based differential analysis is performed (*e.g.*, using **diffHic**), rendering a map of p-values where each p-value $\mathbf{P}_{ij}$ corresponds to the test performed for pixel (i, j). (B) A clustering of the pixels is then obtained, providing K clusters of pixels $\mathcal{C} = \{C_1, \ldots, C_K\}$. The clustering step can take as inputs either the Hi-C matrices — regardless of their condition — or the results of the pixel-level differential analysis. (C) The *post hoc* inference step takes as inputs the p-values computed at step (A) and the clusters of pixels defined at step (B) to compute $\gamma_\alpha(C_k)$, the (minimal) TDP (True Discovery Proportion) for each cluster C_k. (D) An optional step of cluster selection can finally be performed by thresholding $\gamma_\alpha(C_k)$.

3D conformation of the genome between the two conditions $g = 1$ and $g = 2$, *i.e.*, differential clusters of pixels.

2.2. Proposed workflow: hicream

Our approach is implemented in the R package **hicream**, following the workflow illustrated in Figure 1.

(A) A statistical test is performed for each pixel to assess significant differences between conditions, resulting in a p-value map. We chose **diffHic** to perform this differential analysis, as explained in Section 2.2.1;

(B) A set of pixel clusters, $\mathcal{C} = \{C_1, C_2, \cdots, C_K\}$, is either obtained by an automated clustering method, or chosen by the user. Possible clustering choices are discussed in Section 2.2.2, including data-driven methods;

(C) The p-value map and the clusters are combined using a *post hoc* approach described in Section 2.2.3 to provide a lower bound on the true discovery proportion (TDP) in each cluster of $\mathcal{C}$, $(\gamma_\alpha(C_k))_{k=1,\ldots,K}$, with high probability $1 - \alpha$;

(D) (optional) The user can define a TDP threshold to select some of these clusters (those with enough statistical evidence) for further interpretation (*e.g.*, enrichment analyses).

These steps are described in the next sections and further information on the method is also provided in Supplementary Section S1.

2.2.1. Pixel-based differential analysis

Our recent benchmark of pixel-based differential analysis methods [Jorge et al., 2025b] highlighted the wide diversity of available approaches and their heterogeneous statistical performance. Among them, the method implemented in the R/Bioconductor package **diffHic** [Lun and Smyth, 2015] showed strong results and was therefore selected as the default option in **hicream**. For every pixel (i, j), it results in a p-value half-matrix, $(\mathbf{P}_{ij})_{i,\,j=1\ldots,p,\ i\leq j}$, which contains the result of $m \leq \frac{p(p+1)}{2}$ tests performed at the pixel level, where m is the number of pixels for which at least one interaction has been observed in one of the r Hi-C matrices.

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026

2.2.2. Pixel clustering

The second step of the workflow consists of defining pixel clusters, $\mathcal{C} = \{C_1, C_2, \cdots, C_K\}$. Importantly, the **hicream** workflow can cope with any number of pixel clusters, including overlapping ones, without compromising its statistical guarantees. Users may therefore define regions based on prior knowledge or specific assumptions.

In the absence of such prior knowledge or expectations regarding regions of interest, **hicream** proposes an automatic, data-driven clustering method. This strategy identifies clusters of pixels that are similar either in terms of interaction counts (**hicream counts**) or with respect to the results of the differential analysis (**hicream diff**), while respecting the 2D organization of the Hi-C matrix. Both approaches rely on a constrained hierarchical clustering based on Ward's linkage [Randriamihamison et al., 2021]. This method takes as input a matrix $\mathbf{X}$ with m rows. It starts with an initial clustering where each pixel is a cluster and greedily agglomerates pairs of adjacent clusters that provide the smallest change in within-cluster inertia.

In **hicream counts**, for each of the m pixels $q := (i, j)$, $\mathbf{X}$ contains its (normalized by **diffHiC**) counts across replicates and conditions: $\mathbf{X}_{q,\cdot} = (m_{ij}^{1,1}, \ldots, m_{ij}^{1,r_1}, m_{ij}^{2,1}, \ldots, m_{ij}^{2,r_2})$. In **hicream diff**, $\mathbf{X}$ uses the results of the differential analysis, aggregating the log-transformed adjusted p-values and the log-fold change values: $\mathbf{X}_{q,\cdot} = (\log \mathbf{P}_{ij}^{\text{adj}}, \log \text{FC}_{ij})$. The spatial dependence between adjacent pixels from the 2D Hi-C matrix is encoded in the connectivity graph $\mathcal{G}$, where nodes V correspond to pixels and edges link adjacent pixels: Two pixels (i_1, j_1) and (i_2, j_2) are said to be adjacent if $(i_2, j_2) = (i_1, j_1 \pm 1)$ or $(i_2, j_2) = (i_1 \pm 1, j_1)$. Further details on how missing data are handled in the connectivity graph are provided in Supplementary Section S1.

This approach gives a hierarchical clustering that can be represented as a dendrogram and is a collection of clusterings. The final clustering is selected by finding the elbow point of the Ward's linkage.

In practice, we used the algorithm implemented in **scikit-learn** [Pedregosa et al., 2011] which is a greedy algorithm, iteratively merging two clusters while respecting the adjacency constraint. The final clustering was identified using the Python package **kneebow** that implements the elbow criterion. Other clustering strategies and model selection criteria could also be used within our framework.

2.2.3. *Post hoc* inference

Post hoc inference methods [Goeman and Solari, 2011] provide statistical guarantees for arbitrary sets of clusters. For a subset of pixels $C \subset \mathcal{H} := \{1, \ldots, m\}$, let $\text{TDP}(C)$ be the proportion of true positives (or True Discovery Proportion) in C, that is, the (unknown) proportion of pixels (i, j) in C such that the null hypothesis $\text{H}_{0,ij}$ is false. Given a target risk level $\alpha \in (0, 1)$, *post hoc* inference builds, from the data, a function $\gamma_\alpha : \mathcal{H} \to [0, 1]$, such that

$$\mathbb{P}\left(\forall\, C \subset \mathcal{H},\ \text{TDP}(C) \geq \gamma_\alpha(C)\right) \geq 1 - \alpha\,. \quad (1)$$

With probability at least $1 - \alpha$, for any C, the proportion of true positives in C is at least $\gamma_\alpha(C)$. A function γ_α satisfying (1) is called a *post hoc bound* for the TDP.

While both FDR and *post hoc* inference provide statistical control on the proportion of errors ($\text{FDR} = \mathbb{E}(\text{FDP})$ where $\text{FDP} = 1 - \text{TDP}$ is the proportion of false discoveries), we emphasize that they provide fundamentally different statistical guarantees:

- First, while FDR statements are true on average (*in expectation*) over all possible experiments that could have been performed, *post hoc* inference statements are true *with probability* $1 - \alpha$. For example, if for $\alpha = 0.05$ we find $\gamma(C) = 0.875$ for a given cluster C, it means that we are 95% confident that more than 87.5% of the pixels in C are different between the two conditions.
- Second, standard FDR controlling procedures are not applicable to data-driven pixel clusters [Goeman and Solari, 2011]. In contrast, *post hoc* bounds provide statistical guarantees for any number of user-defined subset of hypotheses (here, pixel clusters).

Here, we use the Simes bound introduced by Goeman and Solari [2011] and defined as $\gamma_\alpha : C \mapsto 1 - V_\alpha(C)/|C|$, where

$$V_\alpha(C) = \min_{1 \leq n \leq |C|} \left[\sum_{(i,j) \in C} \mathbb{1}_{\{\mathbf{P}_{ij} > \frac{\alpha n}{m}\}} + n - 1 \right]. \quad (2)$$

The formulation of Equation (2) was obtained by Blanchard et al. [2020] and a linear time implementation is described in Enjalbert-Courrech and Neuvial [2022]. It only requires the availability of a p-value $\mathbf{P}_{ij}$ corresponding to positively dependent (PRDS) test statistics as defined in Benjamini and

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026

Yekutieli [2001]. As such, it is valid under the same assumptions as the widely used BH procedure for FDR control.

2.3. Results evaluation

2.3.1. Semi-synthetic data benchmark

We first evaluated **hicream** using semi-synthetic data with a controlled ground-truth signal, as done previously [Jorge et al., 2025b]. Briefly, real Hi-C contact matrices from the same biological condition (for which no differences are expected) were used as a null hypothesis baseline. Artificial signals were then introduced into specific local regions of a subset of matrices to assess the capacity of the methods to detect them. More specifically, we used five technical replicates of a Hi-C library from a human colon sample [The ENCODE Project Consortium, 2012]. Four of the five technical replicates were used to form two artificial groups of duplicates, while the fifth replicate was used to add counts in predefined regions of the Hi-C matrices, creating known differences between the two groups. This process is illustrated in Supplementary Figure S2.

We tested and compared the two versions of **hicream** and the approach proposed in **diffHic clustering**. Clusters of interest were then selected as:

- **hicream**: clusters for which the *post hoc* bound (for $\alpha = 0.05$) was larger than a given threshold (*i.e.*, as in step (D) of Figure 1);
- **diffHic clustering**: clusters for which the *cluster-level FDR* computed by the package was smaller than a given threshold.

These two threshold values are set independently of each other since the two criteria are not comparable.

For a given threshold and method, we obtained the precision as the proportion of pixels correctly declared differential among all pixels in selected clusters. The recall (power) was also computed similarly. Precision-Recall (PR) curves were obtained by varying the selection threshold from 5% to 100% for **diffHic clustering** and for each unique observed value of γ_α for **hicream**. A perfect classifier would yield a precision of 1 (no false positives) and a recall of 1 (no false negatives). Therefore, the "best" clustering for each method was obtained as the (Precision, Recall) point closest to $(1, 1)$ (which corresponds to two distinct thresholds for the two methods). The evaluation was carried out for three chromosomes (1, 7, and 21) at three resolutions (200 kb, 500 kb, and 1 Mb).

2.3.2. Application to real use case datasets

The first dataset comes from a Hi-C study conducted on a murine cell line under two condition: either CTCF-depleted (CTCF- condition) or control (CTCF+ condition) [Zhang et al., 2021]. Additional data from ChIP-seq experiments against CTCF on the same cell line were also used [Zhang et al., 2019]. Hi-C matrices and ChIP-seq narrow peaks were downloaded from GEO (accessions GSE168251 and GSE129997).

The second real dataset comes from a comparison of mouse cell lines at different stages of neuronal differentiation [Bonev et al., 2017]. Raw Hi-C data from stem cells (ES) and cortical neurons (CN) were downloaded from GEO (accession GSE96107) and processed with the **nfcore/hic** pipeline to obtain four matrices (replicates) of each condition at 50 kb resolution. Corresponding ChIP-seq data were obtained from the UCSC Table Browser.

Each dataset was processed with **hicream diff**. To better characterize the shape of obtained clusters and profile their geometrical diversity with respect to their TDP, we designed three shape-related metrics and assigned them to each cluster: a triangle-, a stripe-, and a square-shape score. Relationships between shape metrics, TDP, and absolute log-fold change (logFC) were characterized using Principal Component Analysis (PCA).

Then, clusters with a minimal TDP greater than 10% across all chromosomes were selected for further analysis.

Clusters selected in the neuronal differentiation study were compared with the results obtained from two other tools designed to identify significant differences in TADs (**TADCompare**, Cresswell and Dozmorov [2020]) or compartments (**dcHiC**, Chakraborty et al. [2022]). Unlike **hicream**, both tools identify 1D location on the genome corresponding to changes. Hence, selected **hicream** clusters were projected onto the genome to identify the corresponding genomic boundaries and their flanking bins. Then, to assess the consistency between **hicream** and **TADCompare**, we tested for the enrichment of these cluster boundaries in differential shifted regions identified by **TADCompare**, compared to flanking bins. As for **dcHiC**, we explored the relationship, at the pixel level, between TDP (related to the cluster in which the pixel is included) and major changes in

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026

compartment conformation (when a pixel is between two "A" regions in condition 1 and between two "B" regions in condition 2, or the opposite).

hicream clusters of both studies were also compared with known biological signal related to genomic organization: We tested for CTCF enrichment at cluster boundaries compared with the flanking bins, by using data on active CTCF binding sites from ChIP-seq experiments. In addition, the distribution of active CTCF binding sites was further characterized by profiling their relative positions with respect to the cluster boundaries.

Finally, cluster boundaries were also tested for enrichment of all known TFBSs from the JASPAR database (CORE Vertebrates 2024) using the AME tool from the MEME Suite (v5.5.4). Last, genes with enriched binding motifs from the TFBS analysis were also subjected to a Gene Ontology functional analysis using the Panther Classification System tool from `https://geneontology.org` (PANTHER GO-Slim Biological Process, Fisher exact test and Bonferroni multiple testing correction).

Further information on data processing and cluster validation is provided in Supplementary Section S2.

2.4. Package implementation and visualization

hicream is implemented as an R package available on CRAN [Jorge et al., 2025a]. Since the default plotting functions provide only static visualizations, we additionally developed an interactive interface based on the R package **animint2** [Sievert et al., 2019]. All results from this study can be explored online at `https://eoojj.github.io/hicream-article-viz/`. A detailed illustration of this interactive visualization platform is provided in Supplementary Figure S3.

3. Results

3.1. On semi-synthetic data, hicream identifies relevant clusters of differential pixels

Figure 2 displays Precision-Recall curves for the three compared approaches (both **hicream** versions and **diffHic clustering**) on the semi-synthetic dataset. **hicream diff** systematically outperforms **diffHiC clustering** and **hicream counts**. **hicream counts** outperforms **diffHic clustering** at 200 kb resolution but does not show better performance at 500 kb or 1 Mb resolution. As the **hicream counts** variant does not outperform the default **hicream diff** method, we chose to focus on **hicream diff** only in the sequel.

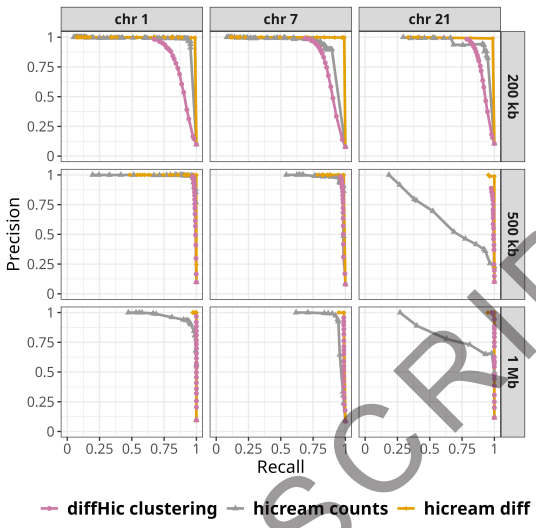


Fig. 2: **PR curves (semi-synthetic data)** computed for three chromosomes of contrasted lengths (chr 1, 7, and 21), at three different bin resolutions (200 kb, 500 kb, and 1 Mb), and for the three methods (both **hicream** versions and **diffHic clustering**). Every point corresponds to a different given threshold value (see Section 2.3.1).

Table 1 gives some characteristics of the "best" selection for each chromosome and resolution for **hicream diff** and **diffHic clustering**. Across all resolutions and chromosomes, **hicream diff** can reach high performance (in terms of precision and recall) with a small number of clusters compared to **diffHic clustering**. Indeed, as can be observed for chromosome 1 at 200 kb, **hicream diff** extracts only 120 clusters – 118 times less than **diffHic clustering** – with a median size of 36 pixels, which is 12 times larger than **diffHic clustering**. Moreover, for chromosome 21 at 1 Mb, despite showing similar performance, both methods largely differ in their output clustering as **hicream diff** summarizes the results in only 7 clusters, while **diffHic clustering** yields 100 clusters, most of them being composed of only one single pixel. Altogether, these results highlight the agreggating and smoothing properties of **hicream**, which allows to properly cluster pixels, facilitating result interpretation.

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026

Table 1. Characteristics of best performing clusterings (semi-synthetic data). The best performance corresponds to the (precision, recall) point closest to $(1, 1)$. For all chromosomes and resolutions and for both **diffHic clustering** and **hicream diff**, information about clustering statistics (number of clusters and their median size in number of pixels) as well as corresponding precision and recall are provided. For **hicream diff**, *post hoc* inference was applied with $\alpha = 0.05$ and the threshold used for cluster selection is given in the column "$\gamma_\alpha(\cdot)$ threshold". Two particular cases are circled in blue and commented in the text.

		diffHic clustering					hicream diff				
Chr.	Res.	Cluster-level FDR	Nb. of clusters	Median cluster size	Precision	Recall	$\gamma_\alpha(\cdot)$ threshold	Nb. of clusters	Median cluster size	Precision	Recall
1	200 kb	0.40	14198	3	0.879	0.802	0.014	120	36.0	0.996	0.989
1	500 kb	0.05	3037	4	0.988	0.971	0.086	72	70.0	0.998	0.999
1	1 Mb	0.05	2634	1	0.967	0.999	0.200	76	10.5	1.000	1.000
7	200 kb	0.45	3994	4	0.904	0.802	0.005	76	33.0	0.997	0.991
7	500 kb	0.05	1131	4	0.987	0.962	0.258	153	10.0	1.000	0.986
7	1 Mb	0.05	1109	1	0.955	0.987	0.698	42	11.0	1.000	0.988
21	200 kb	0.25	314	5	0.939	0.833	0.291	21	34.0	0.989	0.988
21	500 kb	0.05	151	2	0.888	0.975	0.083	19	10.0	0.988	1.000
21	1 Mb	0.10	100	1	0.990	0.990	0.200	7	12.0	1.000	1.000

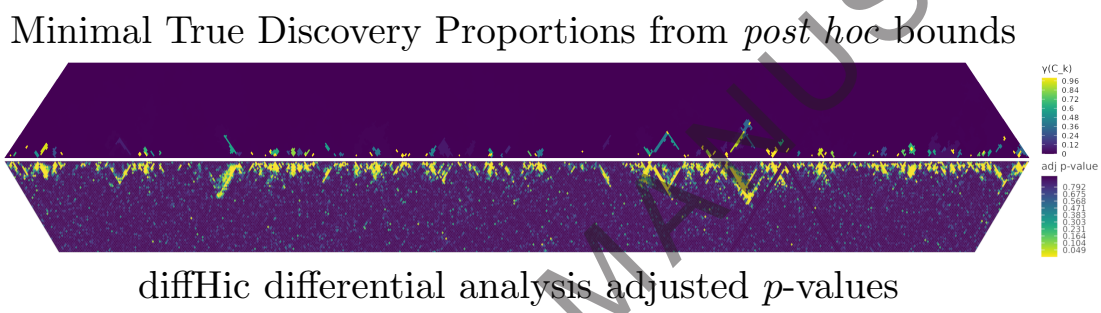


Fig. 3: **Illustration of hicream smoothing and aggregating properties (CTCF depletion study). hicream diff** was applied to chromosome 13 of the CTCF depletion dataset at 200kb resolution. The **kneebow** rule selected 1261 clusters. Upper half: **hicream diff** results. Each cluster of pixels C_k is displayed with a color corresponding to its minimal proportion of true positives $\gamma(C_k)$. Lower half: adjusted p-values obtained from **diffHic** differential analysis. The Hi-C matrix representation has been truncated to focus on the diagonal.

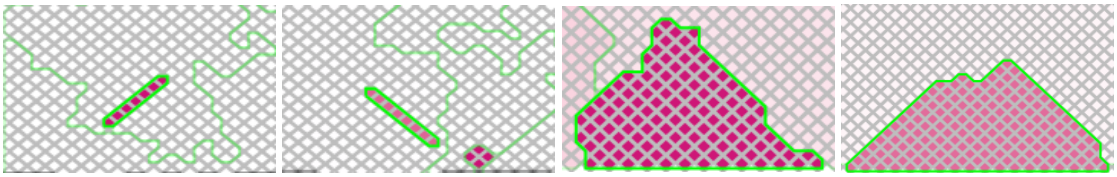
3.2. On real data, **hicream** identifies clear and contrasted clusters of various shapes

By applying **hicream** to real Hi-C datasets, we found that our method reveals sharp clusters from pixel-based results of differential analysis. Figure 3 illustrates these observations in the CTCF depletion data: the upper half displays the cluster-based minimal TDP obtained by **hicream diff**, while the lower half shows the original pixel-based adjusted p-values derived from **diffHic** (chromosome 13, 200 kb). High-TDP clusters in the upper half typically align with groups of pixels exhibiting significant adjusted p-values in the lower half, with sharper and more distinct boundaries. This contrast is improved by the fact that significant yet isolated pixels in the lower half usually get assigned to large clusters with low TDP bounds in the upper half. This illustrates how **hicream** can extend the results of **diffHic**, by removing the background noise, highlighting the most differential regions and providing statistical guarantees.

Importantly, the resulting differential clusters exhibit a wide range of shapes and sizes, showing greater diversity than is typically observed in Hi-C

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026



(a) CTCF depletion study
Chr 1 - Cluster 468
31.2 Mb to 34 Mb
$|C_{468}| = 7$, $\gamma(C_{468}) = 1$
$|\texttt{stripe\_score}| = 1.946$
$q_{C_{468}}^{|\texttt{stripe\_score}|} = 0.999$

(b) CTCF depletion study
Chr 14 - Cluster 310
112.8 Mb to 115.2 Mb
$|C_{310}| = 8$, $\gamma(C_{310}) = 0.75$
$|\texttt{stripe\_score}| = 2.079$
$q_{C_{310}}^{|\texttt{stripe\_score}|} = 0.995$

(c) ND study
Chr 3 - Cluster 60
89.35 Mb to 90.25 Mb
$|C_{60}| = 124$, $\gamma(C_{60}) = 0.99$
$\texttt{triangle\_score} = 0.765$
$q_{C_{60}}^{\texttt{triangle\_score}} = 0.998$

(d) ND study
Chr 6 - Cluster 373
68 Mb to 68.95 Mb
$|C_{373}| = 173$, $\gamma(C_{373}) = 0.72$
$\texttt{triangle\_score} = 0.958$
$q_{C_{373}}^{\texttt{triangle\_score}} = 0.997$

Fig. 4: **Stripe and triangle -shaped differential clusters identified by hicream (CTCF depletion and Neuronal Differentiation studies).** The green lines are cluster boundaries and each cluster C_k is colored according to $\gamma_\alpha(C_k)$. Four selected clusters, highlighted with a slightly more vivid boundary color, are examples of stripe-shaped differential clusters identified with **hicream diff** in the CTCF depletion study at 200 kb resolution, and triangle-shaped differential clusters identified in the neuronal differentiation dataset at 50 kb resolution. For each cluster, information about chromosome, cluster number, cluster position on the chromosome (the left and right positions of the first and last chromosome bins, respectively, among the pixels belonging to this cluster), size (number of pixels), minimal True Discovery Proportion $\gamma_\alpha(C_k)$, shape score (either `stripe_score` or `triangle_score`), and quantile of the shape score for the corresponding cluster is given below the plot.

interaction matrices (Figures 3 and 4). Whereas Hi-C data are generally characterized by a hierarchical TAD-like organization, the focused and precise delineation of groups of differentially proximal interactions makes it possible to highlight the variable components of these canonical structures. In this context, the ability of **hicream** to detect statistically significant clusters of arbitrary shape without imposing *a priori* constraints constitutes a valuable feature.

Supplementary Figure S4 further illustrates the diversity of shapes identified by our method with a PCA of shape metrics, TDP and logFC. It exhibits substantially different patterns across biological datasets. Interestingly, on the CTCF depletion study, high TDPs are associated with large stripe-scores, consistent with the chromatin extrusion process underlying TAD boundary and chromatin loop formation. On the contrary, in the neuronal differentiation study, high TDPs are associated with large triangle-scores, showing that a different biological process (TAD compaction or de-compaction) is at hand in this dataset. Overall, this supports the notion that various patterns of 3D variation drive distinct functional effects, and suggests that **hicream** can capture a broad range of dynamic conformations that are not limited to a specific structure.

To confirm this generality of application, we compared **hicream** clusters with predictions from two structure-level differential methods, both run on the neuronal differentiation data: **TADCompare** to detect variable TAD boundaries, and **dcHiC** to detect differences in chromatin compartmentalization.

Supplementary Figure S5 shows the enrichment of bins identified as differentially shifted by **TADCompare** and enriched in CN (resp. ES) condition at positive (resp. negative) log-fold change cluster boundaries. As a negative control, the same figure shows a depletion of differentially shifted bins enriched in CN (resp. ES) at negative (resp. positive) log-fold change cluster boundaries.

To compare with compartment differences, we extracted pixels found significantly involved in a major compartment change by **dcHiC**: These were pixels "AA2BB" (between two bins in "A" compartment for the ES condition and in "B" compartment for the CN condition) and "BB2AA" (that correspond to the opposite situation). Supplementary Figure S6 shows the distribution of cluster TDP for these specific pixels compared to all other pixels and exhibits a significantly increased TDP for pixels corresponding

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026

to compartment changes (Wilcoxon test, p-values $< 2.22e-16$).

The link between compartment differences and differential clusters identified by **hicream** is not limited to "AA2BB" changes only. Indeed, Supplementary Figure S7 displays a group of 7 differential clusters (TDP$>$ 10%) identified by **hicream** on chromosome 19 of the neuronal differentiation study that represent an "AA2AB" switch. The left projections of the 7 clusters are all located in a region of non-differential bins corresponding to the A compartment in both ES and CN conditions. On the other hand, the right projections of the clusters contain 3 differential bins that switch from A compartment in ES condition to B compartment in CN condition. Therefore, those clusters correspond to pixels representing AA interactions in ES condition and AB interactions in CN condition. This is consistent with the fact that the 7 clusters have a negative log-fold change.

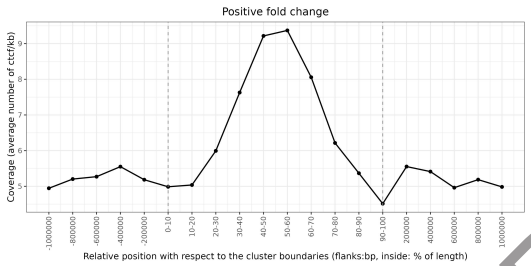
These results confirm that although **hicream** is neither designed nor limited to find differences in TADs or compartments specifically, it is able to identify signal related to multiple levels of chromatin organization at once, showing consistency with structure-level prediction tools.

3.3. Differential clusters identified by **hicream** show relevant TFBS enrichment at their boundaries

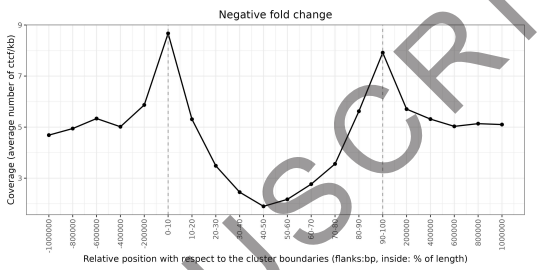
We further investigated the biological relevance of **hicream** clusters by focusing on the boundaries of their genomic projections in both datasets. First, using experimental ChIP-seq data, we observed a significant enrichment of active CTCF binding sites at cluster boundaries compared to their flanking regions, with an average of 21.1 peaks versus 16.8 in the neuronal differentiation study, and 5.7 peaks versus 3.8 in the CTCF depletion study (p-value $= 2.52 \times 10^{-16}$ and 6.26×10^{-8} respectively, Wilcoxon paired test; see also Supplementary Figure S8 for the neuronal differentiation study).

For the CTCF depletion study, we further examined the full CTCF occupancy profiles within and around the projected clusters depending on the sign of their logFC. As shown in Figure 5, the occupancy profiles differ strikingly between clusters with positive and negative logFC values, exhibiting complementary patterns. This difference is consistent with the known effects of CTCF depletion on topologically associating domains (TADs) (see Figure S9 and Supplementary Section S3). In brief, CTCF depletion is known to weaken TAD boundary insulation, allowing increased inter-TAD interactions across boundaries. As a result, pixels with increased interaction values after depletion are expected to localize above and around annotated TAD boundaries in Hi-C matrices, giving rise to clusters with positive logFC values and high CTCF binding site densities. Conversely, as CTCF depletion globally reduces TAD strength, intra-TAD interactions are expected to decrease. This effect generates differential clusters with negative logFC values and TAD-like CTCF profiles at their boundaries.

Last, we extended the enrichment analysis to all known transcription factor binding sites (TFBSs) from the JASPAR database (see Methods). For the neuronal differentiation study, CTCF showed the



(a) Positive log-fold change clusters



(b) Negative log-fold change clusters

Fig. 5: **Distribution of CTCF peak coverage within and around the projected regions of hicream differential clusters at 200 kb resolution (CTCF depletion study).** Positions on the x-axis correspond to left and right flanking regions of the projected boundaries and ten genomic intervals of size 10% of each cluster length between the two boundaries. Top (resp. Bottom): Clusters with a majority of positive (resp. negative) log-fold change pixels.

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026

most significant enrichment (Supplementary File S2). Functional analysis of all enriched transcription factors revealed an overrepresentation of Gene Ontology terms related to cell differentiation and neuronal processes, including regulation of neurogenesis, regulation of cell differentiation, brain development, axon development, and other neuronal functions (Supplementary Table S1).

For the CTCF depletion study, in contrast to the results obtained with the ChIP-seq data, no significantly enriched TFBS were identified. Since ChIP-seq data reflect active CTCF binding sites, whereas TFBS enrichment is based on computational predictions, this discrepancy may suggest that **hicream** more accurately identifies sites that are truly affected by the depletion experiment. Together, these results support the biological relevance of **hicream** differential clusters and demonstrate that **hicream** accurately captures spatial reorganization of chromatin interactions from real Hi-C data.

4. Discussion and conclusion

We have introduced **hicream**, a flexible three-step framework of Hi-C data differential analysis. On a semi-synthetic dataset with controlled ground truth, we compared both versions of **hicream** with the only other method that enables data-driven discovery of differential clusters in Hi-C data, **diffHic** clustering. Across all tested chromosomes and resolutions, both versions of **hicream** outperformed **diffHic clustering**. Fewer but larger differential clusters were obtained, highlighting its ability to capture coherent and interpretable differential genomic structures.

In real case studies, we further showed that **hicream** can detect differential clusters of arbitrary shape, including non-standard yet biologically relevant structures such as stripes. Moreover, the results were consistent with the distribution of active CTCF binding sites and TFBSs at cluster boundaries, further supporting the ability of **hicream** to accurately identify reorganized genomic regions.

Furthermore, the method is highly generic and modular: The method (and implementation) can accommodate alternative choices for differential analysis and clustering than those implemented in **hicream**. In particular, the differential analysis step can be extended to more complex experimental designs by incorporating covariates and testing specific hypotheses through flexible contrasts.

Moreover, because *post hoc* inference provides statistical guarantees independently of the clustering procedure, prior knowledge can also be used to define regions of interest. For example, candidate regions may be derived from structural predictions such as TADs, A/B compartments, or chromatin loops when targeting specific genomic structures. Prior knowledge could also be integrated into the clustering procedure itself, by constraining it to produce clusters of a specific shape (e.g., stripe-like). While the former is straightforward to implement in **hicream** as it only requires specifying candidate regions, the latter would entail designing new clustering procedures able to enforce structural constraints, which constitutes a stimulating methodological perspective for the present work.

Finally, the *post hoc* inference step currently relies on the Simes bound, which may be conservative. In settings with spatial dependence, a calibration strategy has been described to obtain sharper bounds [Blanchard et al., 2020], which is a challenging but interesting direction for future research.

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.

G. Blanchard, P. Neuvial, and E. Roquain. Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics*, 48(3):1281–1303, 2020. doi: 10.1214/19-AOS1847.

B. Bonev, N. M. Cohen, Q. Szabo, L. Fritsch, G. L. Papadopoulos, Y. Lubling, X. Xu, X. Lv, J.-P. Hugnot, A. Tanay, and G. Cavalli. Multiscale 3D genome rewiring during mouse neural development. *Cell*, 171(3):557–572.e24, 2017. doi: 10.1016/j.cell.2017.09.043.

A. Chakraborty, J. G. Wang, and F. Ay. dcHiC detects differential compartments across multiple Hi-C datasets. *Nature Communications*, 13(1):6827, 2022. doi: 10.1038/s41467-022-34626-6.

K. G. Cresswell and M. G. Dozmorov. TADCompare: an R package for differential and temporal analysis

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026

of topologically associated domains. *Frontiers in Genetics*, 11:158, 2020. doi: 10.3389/fgene.2020.00158.

N. Enjalbert-Courrech and P. Neuvial. Powerful and interpretable control of false discoveries in differential expression studies. *Bioinformatics*, 38 (23):5214–5221, 2022. doi: 10.1093/bioinformatics/btac693.

K. Gjoni, L. M. Gunsalus, S. Kuang, E. McArthur, M. Pittman, J. A. Capra, and K. S. Pollard. Comparing chromatin contact maps at scale: methods and insights. *Nature Methods*, 22(4):824–833, Mar. 2025. doi: 10.1038/s41592-025-02630-5.

J. J. Goeman and A. Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4): 584–597, 2011. doi: 10.1214/11-STS356.

D. Hua, M. Gu, X. Zhang, Y. Du, H. Xie, L. Qi, X. Du, Z. Bai, X. Zhu, and D. Tian. DiffDomain enables identification of structurally reorganized topologically associating domains. *Nature Communications*, 15(1):502, 2024. doi: 10.1038/s41467-024-44782-6.

E. Jorge, S. Foissac, T. Hocking, P. Neuvial, and N. Vialaneix. *hicream: HIC diffeREntial Analysis Method*, 2025a. URL `http://CRAN.R-project.org/package=hicream`. R package version 0.0.2 (2025-11-19), published on CRAN. Maintainer. Source code at `https://forge.inrae.fr/scales/hicream`.

E. Jorge, S. Foissac, P. Neuvial, M. Zytnicki, and N. Vialaneix. A comprehensive review and benchmark of differential analysis tools for Hi-C data. *Briefings in Bioinformatics*, 26(2):bbaf074, 2025b. doi: 10.1093/bib/bbaf074.

C. A. Lareau and M. J. Aryee. diffloop: a computational framework for identifying and analyzing differential DNA loops from sequencing data. *Bioinformatics*, 34(4):672–674, 2017. doi: 10.1093/bioinformatics/btx623.

E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009. doi: 10.1126/science.1181369.

A. T. Lun and G. K. Smyth. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, 16:258, 2015. doi: 10.1186/s12859-015-0683-0.

D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015. doi: 10.1016/j.cell.2015.04.004.

E. Maigné, C. Kurylo, M. Zytnicki, and S. Foissac. HiCDOC: chromatin compartment prediction and differential analysis from Hi-C data with replicates. bioRxiv preprint. DOI: 10.1101/2025.09.18.677058, 2025.

R. Mourad. TADreg: a versatile regression framework for TAD identification, differential analysis and rearranged 3D genome prediction. *BMC Bioinformatics*, 23:82, 2022. doi: 10.1186/s12859-022-04614-0.

A. M. Oudelaar, R. A. Beagrie, M. Gosden, S. de Ornellas, E. Georgiades, J. Kerry, D. Hidalgo, J. Carrelha, A. Shivalingam, A. H. El-Sagheer, J. M. Telenius, T. Brown, V. J. Buckle, M. Socolovsky, D. R. Higgs, and J. R. Hughes. Dynamics of the 4d genome during in vivo lineage specification and differentiation. *Nature Communications*, 11(1), June 2020. doi: 10.1038/s41467-020-16598-7.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

N. Randriamihamison, N. Vialaneix, and P. Neuvial. Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints. *Journal of Classification*, 38:363–389, 2021. doi: 10.1007/s00357-020-09377-y.

C. Sievert, S. VanderPlas, J. Cai, K. Ferris, F. U. F. Khan, and T. D. Hocking. Extending ggplot2 for Linked and Animated Web Graphics. *Journal of Computational and Graphical Statistics*, 28(2):299–308, 2019. doi: 10.1080/10618600.2018.1513367.

M. Spielmann, D. G. Lupiáñez, and S. Mundlos. Structural variation in the 3D genome. *Nature*

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026

Reviews Genetics, 19(7):453–467, 2018. doi: 10.1038/s41576-018-0007-0.

The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. doi: 10.1038/nature11247.

H. Zhang, D. J. Emerson, T. G. Gilgenast, K. R. Titus, Y. Lan, P. Huang, D. Zhang, H. Wang, C. A. Keller, B. Giardine, R. C. Hardison, J. E. Phillips-Cremins, and G. A. Blobel. Chromatin structure dynamics during the mitosis-to-G1 phase transition. *Nature*, 576(7785):158–162, 2019. doi: 10.1038/s41586-019-1778-y.

H. Zhang, J. Lam, D. Zhang, Y. Lan, M. W. Vermunt, C. A. Keller, B. Giardine, R. C. Hardison, and G. A. Blobel. CTCF and transcription influence chromatin structure re-configuration after mitosis. *Nature Communications*, 12:5157, 2021. doi: 10.1038/s41467-021-25418-5.

5. Data and scripts availability

All datasets used in this article are available or described at `https://doi.org/10.57745/IEUSLV`. The scripts reproducing the results of this manuscript are all provided at `https://forge.inrae.fr/scales/replication-repository-for-hicream-article`.

6. Competing interests

No competing interest is declared.

7. Author contributions statement

PN, NV, and SF designed the study. All authors conducted the research and analyzed the data. EJ, PN, NV, and SF wrote the initial draft. All authors have read and approved the final manuscript.

8. Acknowledgments

We would like to thank the Genotoul-Bioinfo platform (Toulouse Occitanie) for providing computational resources and assistance.

9. Funding

This work was supported by the INRAE DIGIT-BIO program and by a Toulouse INP-ENSAT fellowship.

ACCEPTED MANUSCRIPT

Downloaded from academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btag631/8770358 by Université de Sherbrooke user on 03 September 2026